Computational social science

Data & Digital Traces

About you

Which type of degree are you pursuing?
Bachelor’s or Masters?

About you

Which degree program are you enrolled in?
Complex systems, Information networks,
CCIS, or Others?

About you

Do you have any background/training/ special interest in the social sciences?

About me

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

Ph.D. in Computer Science

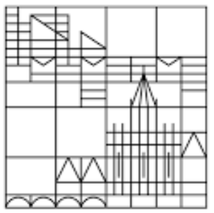
LEIBNIZ INSTITUTE
FOR MEDIA RESEARCH
HANS-BREDOW-INSTITUT

Visiting postdoctoral fellow
“Algorithmed Public Spheres”

gesis Leibniz Institute for the Social Sciences

Postdoctoral researcher
CSS Dept.

Universität Konstanz



Assistant Professor of CSS
Centre for Data & Methods

A! **Aalto University**

Assistant Professor
Department of Comp. Sc.

My research

Studying our internet mediated lives

My research

Studying our internet mediated lives

Combine **digital behavioral data** with **surveys** and **experiments** to study **human behavior** on the web and its impact on individuals' **opinions, attitudes and behaviors**

My research

Studying our internet mediated lives

Combine **digital behavioral data** with **surveys** and **experiments** to study **human behavior** on the web and its impact on individuals' **opinions, attitudes and behaviors**

Course structure

Period IV

Week	Lecture	Exer. dl	Ext. dl	Topic
1	~~Feb 27~~	~~Mar 3~~	Mar 15	~~Introduction to CSS~~
2	~~Mar 6~~	~~Mar 10~~	Mar 22	~~Artificial societies & agent-based models~~
3	Mar 13	Mar 17	Mar 29	Data & digital traces
4	Mar 20	Mar 24	Apr 5	Counting things & analysing text
5	Mar 27	Mar 31	Apr 12	Social networks: structure
6	Apr 3	*	-	Introduction to the project

***Project deadline: May 26**
Project peer review: June 2

Period V

Week	Lecture	Exercise dl	Ext. dl	Topic
7	Apr 24	May 5	May 10	Ethics, privacy, legal
-	-	-	-	WAPPU
8	May 8	May 12**	May 24	Agent-based models & emergence
9	May 15	May 19***	May 31	Social networks: dynamics
10	May 22	May 26***	June 7	Experiments & interventions at scale
11	May 29	-	-	Computing for social good

****Bonus round**
*****Only lecture questions**

SOCIAL SCIENCE

Computational Social Science

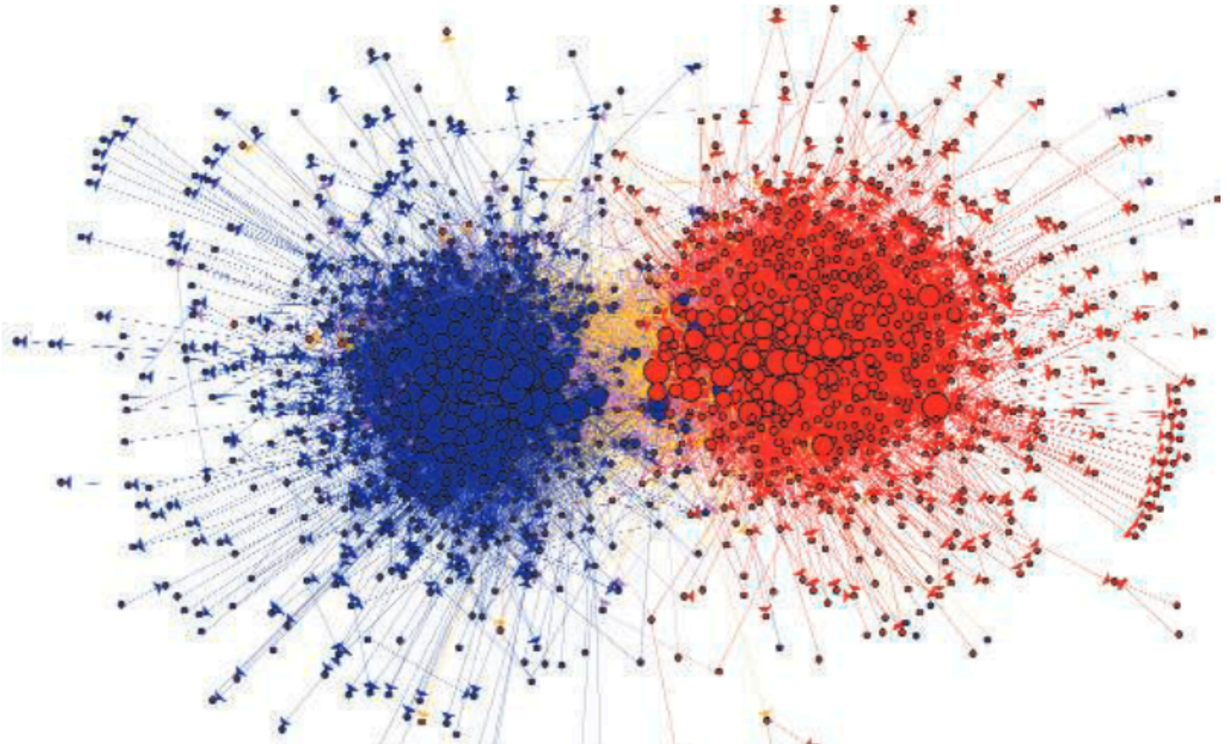
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the



[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University

SOCIAL SCIENCE

Computational Social Science

David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tor[illegible]

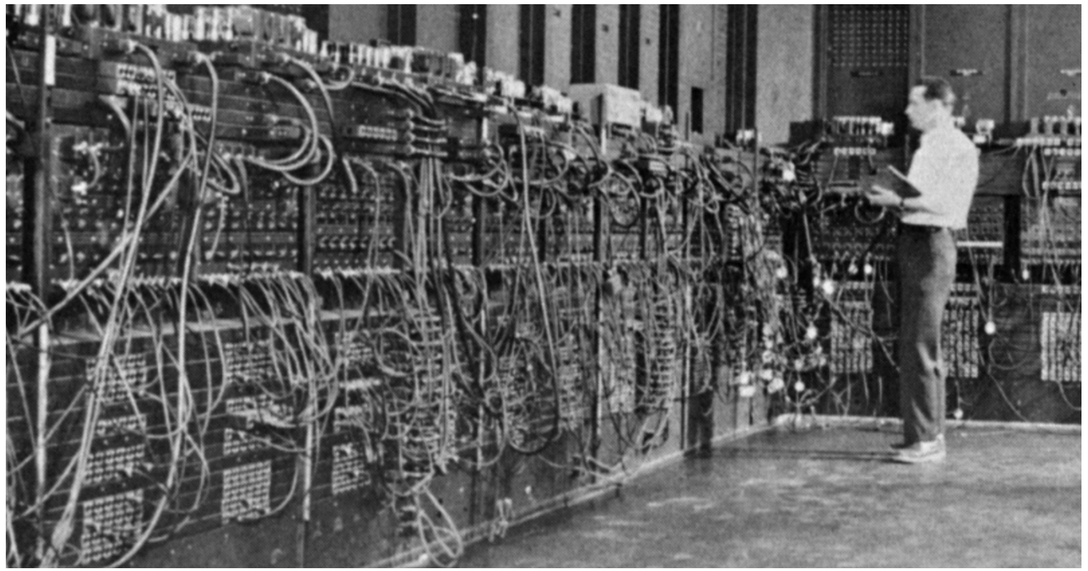
A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviours.

been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in govern-

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University

Amount of data grows exponentially



- The Large Hadron Collider at CERN produced 40 zettabytes (10^{21}) of data
- Amazon Web Services store 500 exabytes (10^{18}) of data

https://arxiv.org/abs/2202.07659

Big data in CSS

Big data in CSS

What is big data?

When is it useful?

https://www.bitbybitbook.com/en/1st-ed/

What is big data in CSS?

- Behavioral data—who does what and when
- Analog age - rare and expensive
- Digital age - plentiful, recorded, stored and analyzable

What is big data in CSS?

- Data about behavior—who does what and when
- Analog age - rare and expensive
- Digital age - plentiful, recorded, stored and analyzable

Give some examples of you generating digital behavioral data in your daily life.

What is big data in CSS?

- **Digital traces**: digital behavioral data produced as side product of other activities

  -> data not created for the purpose of research
- Observational data or “found” data
- Fundamentally different from conventional social science data: questionaires, experiments, …

Example: Facebook social network

- Facebook graph in 2011:
  - 721 million active users (~10% of global population)
  - 69 billion friendship connections (190 friends on average)

J. Ugander, B. Karrer, L. Backstrom, C. Marlow "The Anatomy of the Facebook Social Graph", https://arxiv.org/abs/1111.4503 (2011)

Example: Facebook social network

- Facebook graph in 2011:
  - 721 million active users (~10% of global population)
  - 69 billion friendship connections (190 friends on average)

Heard of “six degree of separation” or small world theory or Milgram’s experiment?

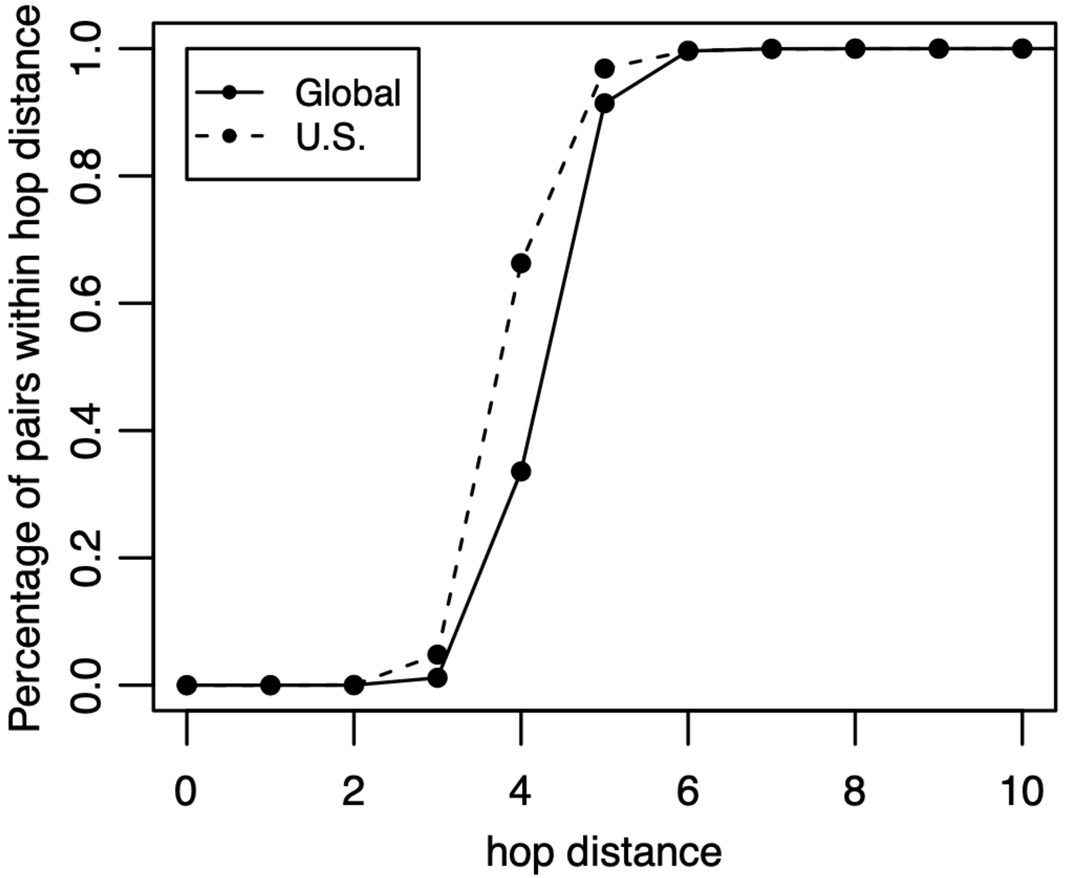
J. Ugander, B. Karrer, L. Backstrom, C. Marlow "The Anatomy of the Facebook Social Graph”, https://arxiv.org/abs/1111.4503 (2011)

Example: Facebook social network

- Facebook graph in 2011:
  - 721 million active users (~10% of global population)
  - 69 billion friendship connections (190 friends on average)
- “Six degrees of separation” (small world hypothesis)

  -> 5.2 in Milgram’s experiments in US in 60s

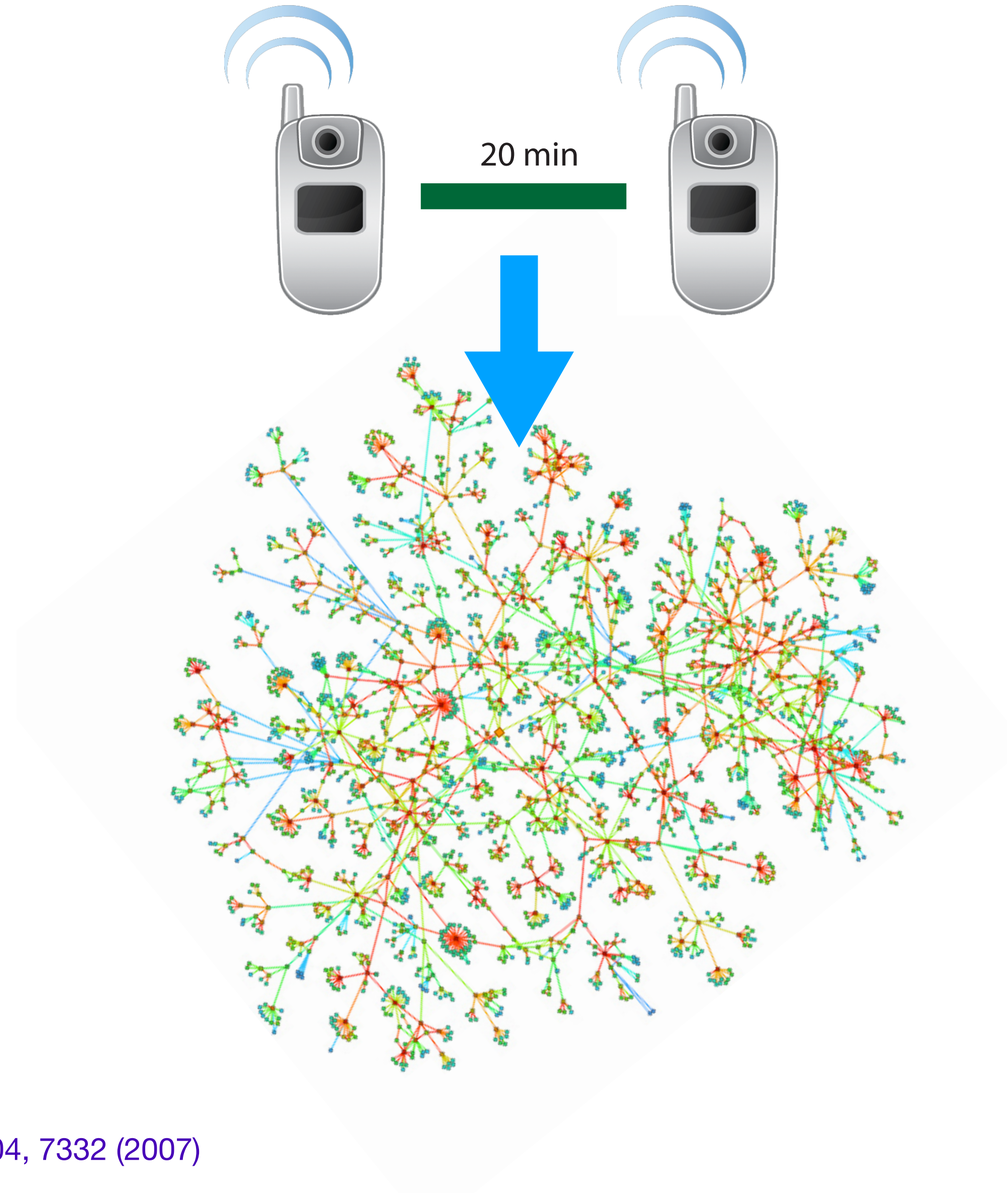
  -> 4.7 in Facebook (in 2011)



J. Ugander, B. Karrer, L. Backstrom, C. Marlow "The Anatomy of the Facebook Social Graph”, https://arxiv.org/abs/1111.4503 (2011)

Example: Call data records

- Social network built out of data on call durations of customers of a mobille phone operator, 7 million customers, 18 months, 300 million calls



Onnela, et al., Proc. Natl. Acad. Sci. (USA) 104, 7332 (2007)

Example: Call data records

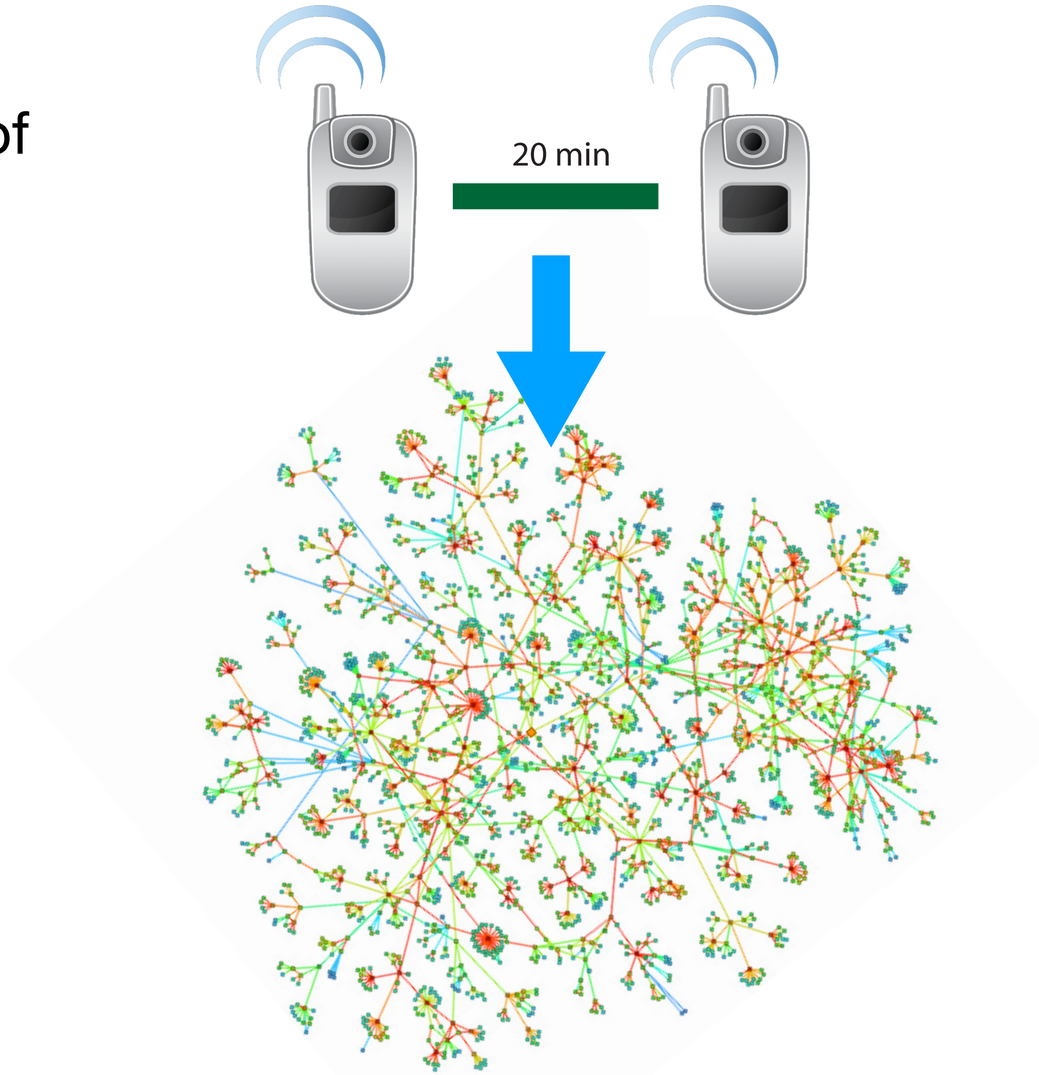
- Social network built out of data on call durations of customers of a mobille phone operator, 7 million customers, 18 months, 300 million calls

Heard of Granovetter’s hypothesis or “the strength of weak ties”?

Onnela, et al., Proc. Natl. Acad. Sci. (USA) 104, 7332 (2007)

Example: Call data records

- Social network built out of data on call durations of customers of a mobille phone operator, 7 million customers, 18 months, 300 million calls
- “Granovetter’s hypothesis” in 1973
  - strong ties within social groups, weak ties between groups
  - weak ties are especially pivotal in the flow of information
- Mobile data proved the hypothesis at a scale of a country



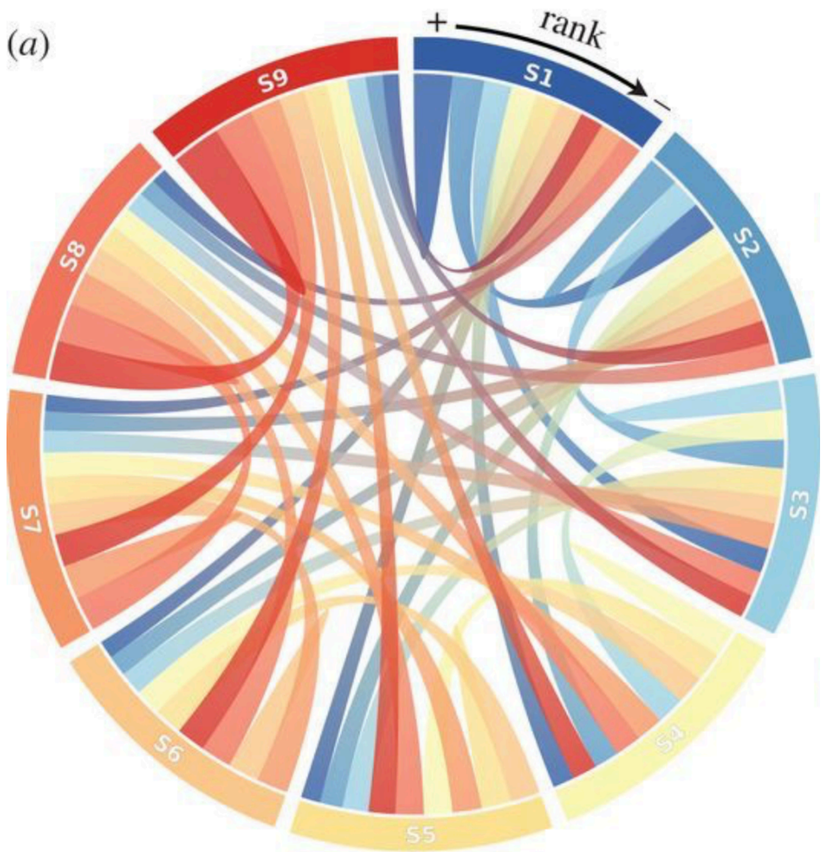
Onnela, et al., Proc. Natl. Acad. Sci. (USA) 104, 7332 (2007)

Example: phone calls + credit data

- ~8 billion calls and sms between 112 million mobile phone customers in Mexico

- Combined with banking data (purchases, loans etc.) of 6 million customers

Map of socioeconomic stratification:



Leo, Y., Fleury, E., Alvarez-Hamelin, J. I., Sarraute, C., & Karsai, M. (2016). Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, *13*(125), 20160598.

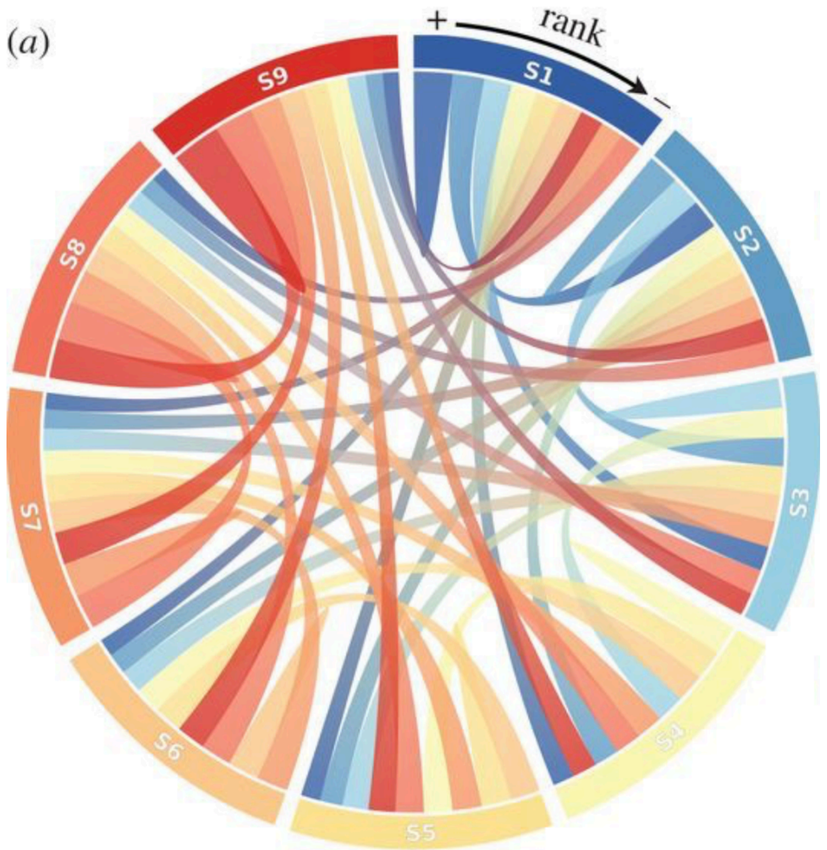
Example: phone calls + credit data

- ~8 billion calls and sms between 112 million mobile phone customers in Mexico

- Combined with banking data (purchases, loans etc.) of 6 million customers

- Studied socio-economic stratification in the Mexican society
  - people better connected to their own socioeconomic class (and live closer) rather than to other classes

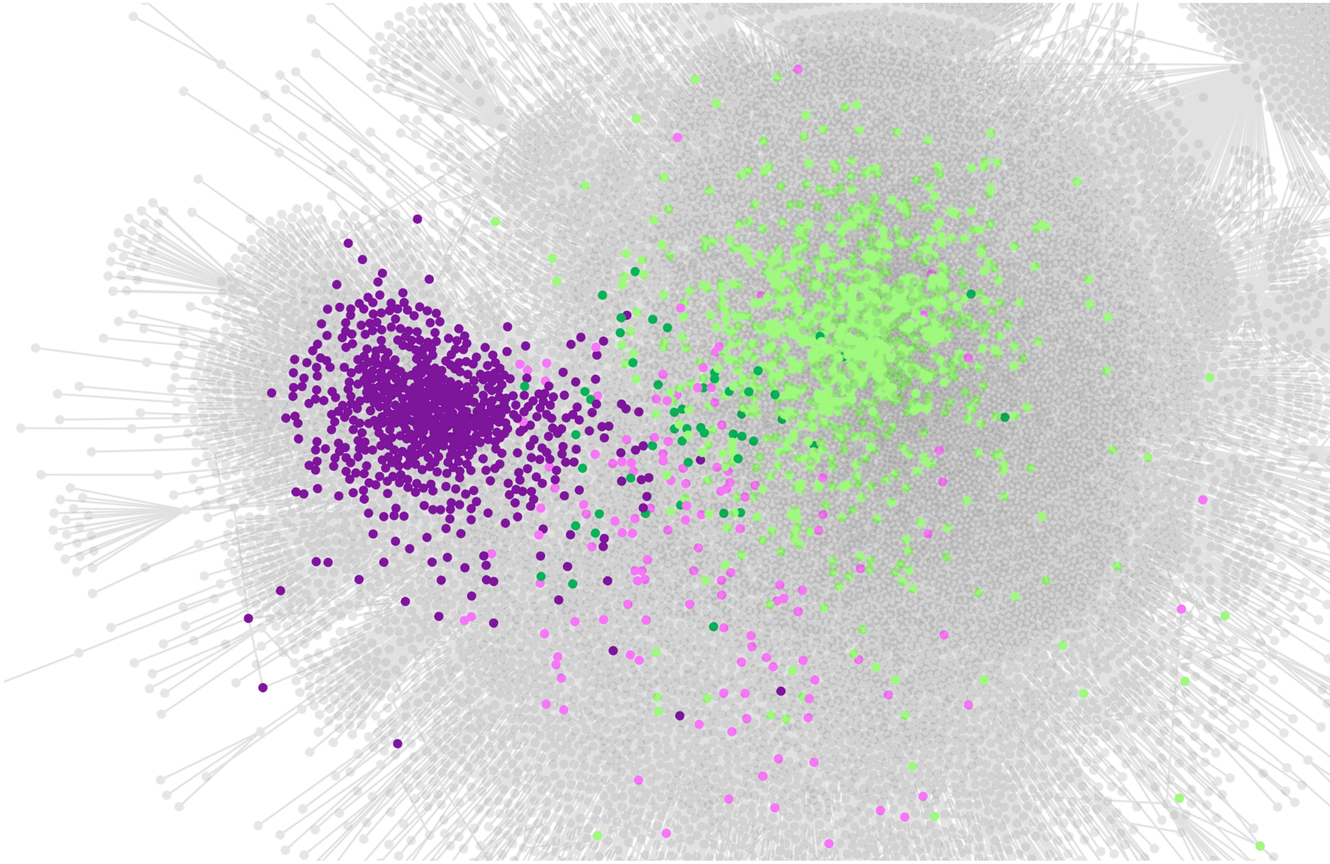
Map of socioeconomic stratification:



Leo, Y., Fleury, E., Alvarez-Hamelin, J. I., Sarraute, C., & Karsai, M. (2016). Socioeconomic correlations and stratification in social-communication networks. *Journal of The Royal Society Interface*, *13*(125), 20160598.

Example: Twitter data

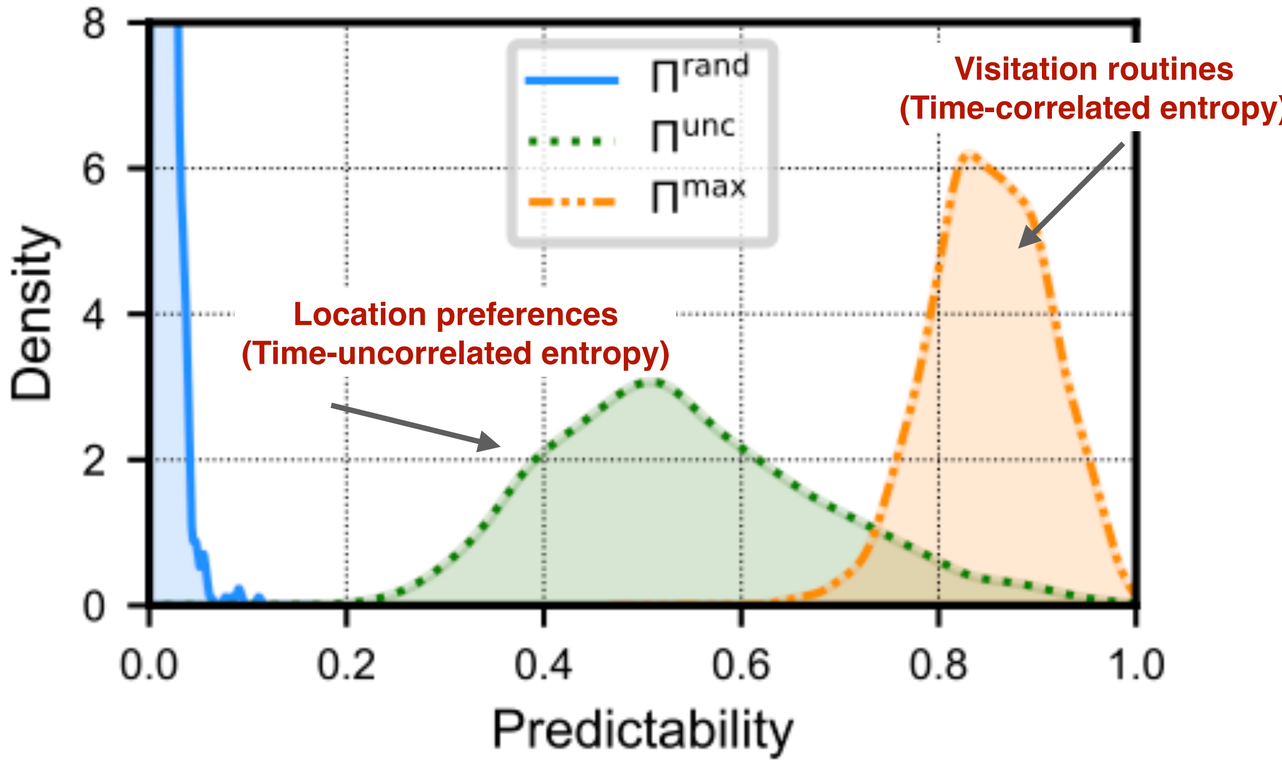
- Twitter has around 200 million daily users
- Academic access: 10 million tweets per month (!)
- Figure: political discussion network in Finland during the 2019 elections (coloured nodes: climate and immigration)
- Alignment between climate and immigration politics is high, with accounts sorted into clear groups



THY Chen, A Salloum, A Gronow, T Ylä-Anttila, M Kivelä "Polarization of Climate Politics Results from Partisan Sorting: Evidence from Finnish Twittersphere." Global Environmental Change 71 (2021): 102348.

Example: Web browsing traces

- Detailed URL-level browsing traces for 2148 German internet users, 9 million web visits to 50 thousand unique domains
- Individuals exhibit routineness while browsing the web
- These routines or repetitive patterns make their browsing behavior predictable to 85% on average



Kulshrestha, Juhi, et al. "Web routineness and limits of predictability: Investigating demographic and behavioral differences using web tracking data." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021.

Big data in CSS

What is big data?

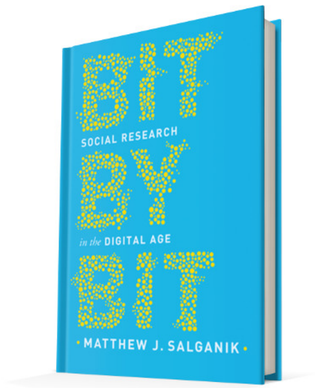
When is it useful?

https://www.bitbybitbook.com/en/1st-ed/

When is big data useful?

10 characteristics of big data

1. Big
2. Always on
3. Nonreactive
4. Incomplete
5. Inaccessible
6. Nonrepresentative
7. Drifting
8. Algorithmically confounded
9. Dirty
10. Sensitive

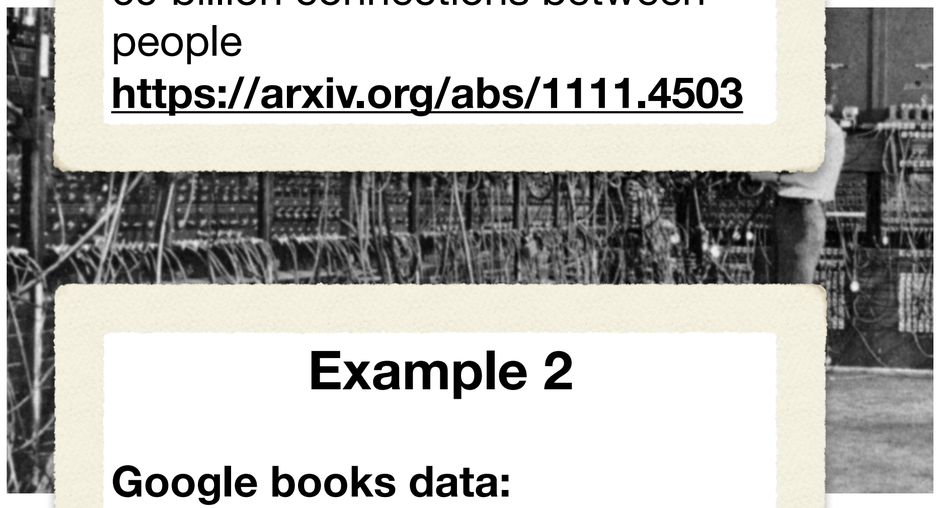


1. Big

- Data can be massive (billions of data points)

Example 1

Facebook data from 2011:
721 million people
69 billion connections between people
https://arxiv.org/abs/1111.4503



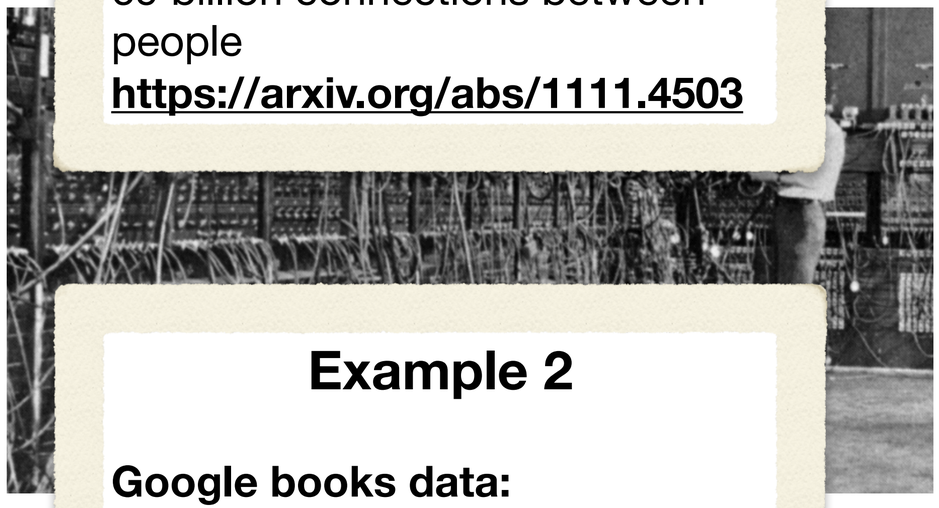
Example 2

Google books data:
5 million books, 4% of all books
500 billion words
https://doi.org/10.1126/science.1199644

1. Big

Example 1

Facebook data from 2011:
721 million people
69 billion connections between people
https://arxiv.org/abs/1111.4503



Example 2

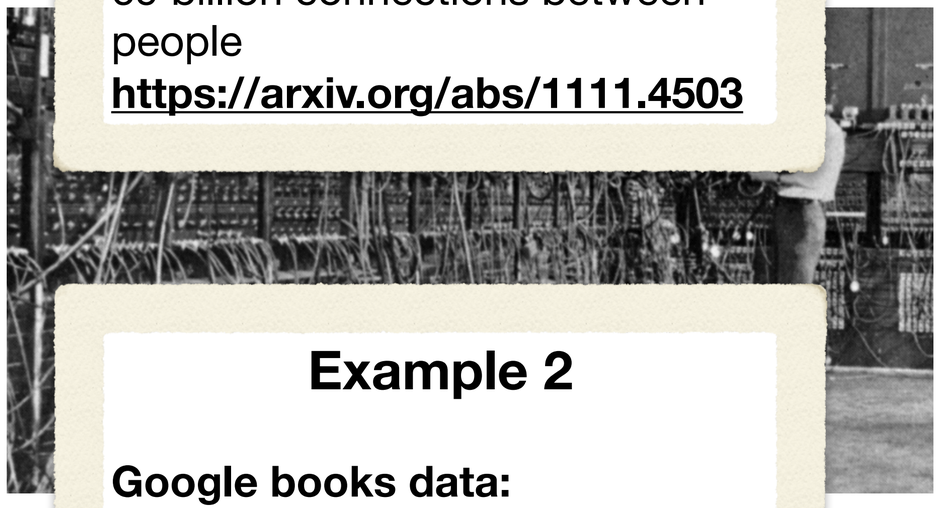
Google books data:
5 million books, 4% of all books
500 billion words
https://doi.org/10.1126/science.1199644

- Data can be massive (billions of data points)
- Size not a value in itself. Ask specifically, how is it useful?
  - Studying rare event, e.g., changes in rate of irregular verb conjugation with Google books data
  - Studying heterogeneity, provides enough data points to statistically test hypotheses e.g., women's browsing more predictable than men's
  - Detecting small differences, e.g., selecting between two public health interventions, one slightly more effective => multitude of lives saved

1. Big

Example 1

Facebook data from 2011:
721 million people
69 billion connections between people
https://arxiv.org/abs/1111.4503



Example 2

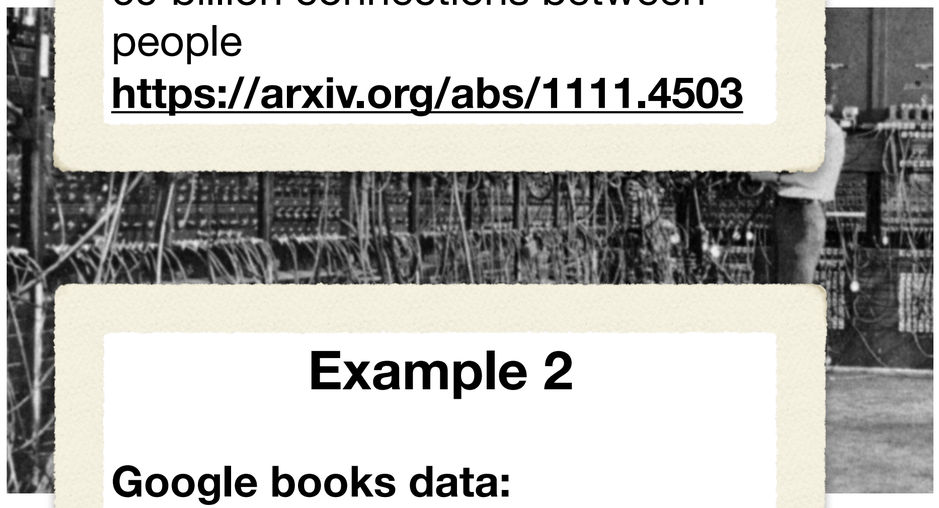
Google books data:
5 million books, 4% of all books
500 billion words
https://doi.org/10.1126/science.1199644

- Data can be massive (billions of data points)
- Size not a value in itself. Ask specifically, how is it useful?
  - Studying rare event, e.g., changes in rate of irregular verb conjugation with Google books data
  - Studying heterogeneity, provides enough data points to statistically test hypotheses e.g., women's browsing more predictable than men's
  - Detecting small differences, e.g., selecting between two public health interventions, one slightly more effective => multitude of lives saved
- Can distract from other problems with the data, e.g., systematic errors from biases in data creation

1. Big

Example 1

Facebook data from 2011:
721 million people
69 billion connections between people
https://arxiv.org/abs/1111.4503



Example 2

Google books data:
5 million books, 4% of all books
500 billion words
https://doi.org/10.1126/science.1199644

- Data can be massive (billions of data points)
- Size not a value in itself. Ask specifically, how is it useful?
  - Studying rare event, e.g., changes in rate of irregular verb conjugation with Google books data
  - Studying heterogeneity, provides enough data points to statistically test hypotheses e.g., women's browsing more predictable than men's
  - Detecting small differences, e.g., selecting between two public health interventions, one slightly more effective => multitude of lives saved
- Can distract from other problems with the data, e.g., systematic errors from biases in data creation

Conventional data:
Sample sizes small -> statistical significance and random fluctuations a major problem

2. Always on

- Data collected continuously, not at specific times

  -> Longitudinal studies

  -> Study of unexpected events

  -> Produce real-time estimates

2. Always on

- Data collected continuously, not at specific times

  -> Longitudinal studies

  -> Study of unexpected events

  -> Produce real-time estimates

Example

Twitter data on Ukrainian war:
Millions of tweets globally about the discussion on Ukrainian war exactly when it started
https://arxiv.org/abs/2203.07488

2. Always on

- Data collected continuously, not at specific times

  -> Longitudinal studies

  -> Study of unexpected events

  -> Produce real-time estimates

Example

Twitter data on Ukrainian war:
Millions of tweets globally about the discussion on Ukrainian war exactly when it started
https://arxiv.org/abs/2203.07488

Conventional data:
Collected at predetermined times -> Longitudinal studies are expensive; may miss unexpected events

3. Nonreactive

- People can change their behaviour when they know they are being observed by researchers (reactivity)

3. Nonreactive

- People can change their behaviour when they know they are being observed by researchers (reactivity)
- Collected in the background, without people paying attention to the collection

  -> The study doesn’t change the behaviour

Example 1

Mobility data collected by phone operators:
People are tracked by operators without them realising it (data is aggregated)

3. Nonreactive

- People can change their behaviour when they know they are being observed by researchers (reactivity)
- Collected in the background, without people paying attention to the collection

  -> The study doesn't change the behaviour

  -> Might not be "unfiltered" behaviour, but not because of attention of researchers (social desirability bias)

Example 1

Mobility data collected by phone operators:
People are tracked by operators without them realising it (data is aggregated)

Example 2

Social media data:
Data on discussions on Facebook don't represent private discussions, awareness that they can be seen by others

3. Nonreactive

- People can change their behaviour when they know they are being observed by researchers (reactivity)
- Collected in the background, without people paying attention to the collection

  -> The study doesn't change the behaviour

  -> Might not be "unfiltered" behaviour, but not because of attention of researchers (social desirability bias)

Conventional data:
People know they are being studied-> Change of behavior

Example 1

Mobility data collected by phone operators:
People are tracked by operators without them realising it (data is aggregated)

Example 2

Social media data:
Data on discussions on Facebook don't represent private discussions, awareness that they can be seen by others

4. Incomplete

- Data collection not designed for the research purpose -> misses important data

  1. Demographics
  2. Behavior on other platforms
  3. Data to operationalise *theoretical constructs* (construct validity)

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

4. Incomplete

- Data collection not designed for the purpose -> misses important data

  1. Demographics
  2. Behavior on other platforms
  3. Data to operationalise *theoretical constructs* (construct validity)

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

Can you think of what incompleteness-related challenges you would encounter with this data to study the association between social relationship and phone calls?

4. Incomplete

- Data collection not designed for the purpose -> misses important data
  1. Demographics
  2. Behavior on other platforms
  3. Data to operationalise *theoretical constructs* (construct validity)

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

Example, continued

- No access to reason or content of the phone call
- Limited access to demographics of the people
- No access to calls made via WhatsApp etc.

4. Incomplete

- Data collection not designed for the purpose -> misses important data
  1. Demographics
  2. Behavior on other platforms
  3. Data to operationalise *theoretical constructs* (construct validity)

Conventional data:
Designed to collect data exactly answering the research question

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

Example, continued

- No access to reason or content of the phone call
- Limited access to demographics of the people
- No access to calls made via WhatsApp etc.

5. Inaccessible

- Data owned by companies and governments, difficult to access
  - Data is valuable

Example 1

Phone calls & mobility from operator data:
Data owned by mobile phone companies, sold commercially

5. Inaccessible

- Data owned by companies and governments, difficult to access
  - Data is valuable
  - Privacy, legal, ethical problems (more in week 7)

Example 1

Phone calls & mobility from operator data:
Data owned by mobile phone companies, sold commercially

Example 2

Social media data:
Data owned by companies, limited access to researchers, resharing difficult (Meta, Twitter etc)

5. Inaccessible

- Data owned by companies and governments, difficult to access
  - Data is valuable
  - Privacy, legal, ethical problems (more in week 7)

Conventional data:
Data collected and owned by the researchers

Example 1

Phone calls & mobility from operator data:
Data owned by mobile phone companies, sold commercially

Example 2

Social media data:
Data owned by companies, limited access to researchers, resharing difficult (Meta, Twitter etc)

6. Nonrepresentative

- People in data are not uniformly randomly sampled from some larger population of interest

  -> Results might not generalise outside of the sample, especially to the population from which the sample was drawn => out-of-sample generalisations may not work

  -> Within-sample comparisons work

6. Nonrepresentative

- People in data are not uniformly randomly sampled from some larger population of interest

  -> Results might not generalise outside of the sample, especially to the population from which the sample was drawn => out-of-sample generalisations may not work

  -> Within-sample comparisons work

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

6. Nonrepresentative

- People in data are not uniformly randomly sampled from some larger population of interest

  -> Results might not generalise outside of the sample, especially to the population from which the sample was drawn => out-of-sample generalisations may not work

  -> Within-sample comparisons work

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

Can you think of what nonrepresentativeness-related challenges you would encounter with this data to study the association between social relationship and phone calls?

6. Nonrepresentative

- People in data are not uniformly randomly sampled from some larger population of interest

  -> Results might not generalise outside of the sample, especially to the population from which the sample was drawn => out-of-sample generalisations may not work

  -> Within-sample comparisons work

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

Example continued

Phone calls to social relationships:
People and connections with biased sampling:
- Demographics of people using phone calls
- Types of relationships captured by phone calls
- Only within company calls

6. Nonrepresentative

- People in data are not uniformly randomly sampled from some larger population of interest

  -> Results might not generalise outside of the sample, especially to the population from which the sample was drawn => out-of-sample generalisations may not work

  -> Within-sample comparisons work

Conventional data:
Researcher can define the population and design a sampling procedure

Example

Phone calls to social relationships:
Theoretical construct: social relationship, tie strength
Operationalisation: Number or phone calls between 2 people (from operation billing records)
https://doi.org/10.1140/epjds/s13688-020-00256-5

Example continued

Phone calls to social relationships:
People and connections with biased sampling:
- Demographics of people using phone calls
- Types of relationships captured by phone calls
- Only within company calls

7. Drifting

- Longitudinal changes because big data systems change all the time:
  - *population drift*: who is using them changes
  - *behavioral drift*: how people are using the system changes
  - *system drift*: the system itself changes

7. Drifting

- Longitudinal changes because big data systems change all the time:
  - *population drift*: who is using them changes
  - *behavioral drift*: how people are using the system changes
  - *system drift*: the system itself changes

Example

Phone calls to social relationships (longitudinal study):
Population drift: demographics of phone owners changed
Behavioral drift: people switched from phone calls to messaging apps
System drift: The way operator records data changes

7. Drifting

- Longitudinal changes because big data systems change all the time:
  - *population drift*: who is using them changes
  - *behavioral drift*: how people are using the system changes
  - *system drift*: the system itself changes

Example

Phone calls to social relationships (longitudinal study):
Population drift: demographics of phone owners changed
Behavioral drift: people switched from phone calls to messaging apps
System drift: The way operator records data changes

Conventional data:
Researchers try to make measurements as stable as possible

8. Algorithmically confounded

- Systems are guiding the users to specific behavior

  -> observed behavior might be due to the design of the system, not ‘natural’ behavior

Example 1

Number of friends on Facebook:
Anomalously high number of users with 20 friends
FB algorithm recommends friends for new users till they reach 20 friends
https://doi.org/10.48550/arXiv.1111.4503

8. Algorithmically confounded

- Systems are guiding the users to specific behavior

  -> observed behavior might be due to the design of the system, not ‘natural’ behavior

  -> designers of a system use social theories to design their sytem’s behavior => performativity

Example 1

Number of friends on Facebook:
Anomalously high number of users with 20 friends
FB algorithm recommends friends for new users till they reach 20 friends
https://doi.org/10.48550/arXiv.1111.4503

Example 2

Triadic closure in online social networks:
Social theory: if A know B and C, then B and C more likely to know each other than random pair of people
Algorithm: Recommends people you might know, exactly based on this rule

8. Algorithmically confounded

- Systems are guiding the users to specific behavior

  -> observed behavior might be due to the design of the system, not ‘natural’ behavior

  -> designers of a system use social theories to design their sytem’s behavior => performativity

Example 1

Number of friends on Facebook:
Anomalously high number of users with 20 friends
FB algorithm recommends friends for new users till they reach 20 friends
https://doi.org/10.48550/arXiv.1111.4503

Ex

Triadic closure in on
Social theory: if A kn
more likely to know e

Can you think of examples of algorithmic confounding you may have encountered?

9. Dirty

- Includes data that is of not interest and might distort the results, not real social behavior (e.g., bots)

  -> data cleaning

Example 1

Online social networks:
Bots, and other automated content

9. Dirty

- Includes data that is of not interest and might distort the results, not real social behavior (e.g., bots)

  -> data cleaning

  -> what is dirty can differ based on research question

Example 1

Online social networks:
Bots, and other automated content

Example 2

Wikipedia edits:
Human edits vs. Automated bot edits
Only human edits - how humans contribute to wikipedia
All edits - studying wikipedia as an ecosystem

9. Dirty

- Includes data that is of not interest and might distort the results, not real social behavior (e.g., bots)

  -> data cleaning

  -> what is dirty can differ based on research question

Example 1

Online social networks:
Bots, and other automated content

Example 2

Wikipedia edits:
Human edits vs. Automated bot edits
Only human edits - how humans contribute to wikipedia
All edits - studying wikipedia as an ecosystem

Conventional data:
Needs some cleaning but not as big issue as big data

10. Sensitive

- Detailed personal data often sensitive
- Even harmless looking data might be sensitive for small number of people
- People might not realise their data is being used => no consent => ethical concerns

Conventional data:
Researchers can ask for informed consent & collect exactly the data that is needed

Summary

- Big social data produced by *digital traces*
- Different from conventional social science data

  -> generally good: big, always-on, and nonreactive

  -> generally bad: incomplete, inaccessible, nonrepresentative, drifting, algorithmically confounded, inaccessible, dirty, and sensitive